PATENT ABSTRACTS OF JAPAN

(11)Publication number : **04-191977**
(43)Date of publication of application : **10.07.1992**

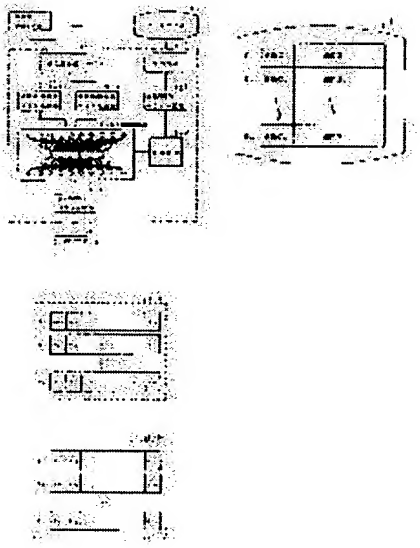
(51)Int.Cl. **G06F 15/40**
G06F 15/18

(21)Application number : **02-323540**
(22)Date of filing : **27.11.1990**

(71)Applicant : **NIPPON TELEGR & TELEPH CORP <NTT>**
(72)Inventor : **KUWABARA SATOSHI**

(54) SIMILARITY CORRECTION PROCESSING METHOD

(57)Abstract:
PURPOSE: To find out a program with a high utility value by retrieving similar data reflecting an experienced retrieval tendency when a characteristic with a constitution is turned to be a retrieval key and the similarity retrieval of the similar data is performed based on a data base.
CONSTITUTION: Routing Sti is calculated between retrieval data Ct and the whole characteristic data Ci (i-1 to N) in the data base by a similarity genera tion part 3-1, one N-dimensional similarity vector St=(st1,st2,...stN) is prepared to be stored in similarity vector storage part 3-3 for retrieval. Corresponding rank vector Et= (et1,et2,...etN) is outputted to a rank vector storage part 3-8 for retrieval by inputting this similarity vector in a learned neural circuit net work 3-6. In this rank vector, the rank to be displayed indicates the rank of a similar data pair to be retrieved actually, and when the learning in the former stage is sufficiently refined, this rank reflects the experienced retrieval tendency. Thus, the program with high utility value can be find out.



LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

Copyright (C); 1998,2003 Japan Patent Office

⑲日本国特許庁(JP)　⑪特許出願公開

⑫公開特許公報(A)　平4−191977

⑤Int.Cl.⁵	識別記号	庁内整理番号	㊸公開　平成4年(1992)7月10日
G 06 F 15/40	510 J	7056−5L	
15/18		8945−5L	

審査請求　未請求　請求項の数 1　(全9頁)

㊹発明の名称　類似度補正処理方法

㉑特　願　平2−323540

㉒出　願　平2(1990)11月27日

㉗発明者　桑原　敏　東京都千代田区内幸町1丁目1番6号　日本電信電話株式会社内

㉑出願人　日本電信電話株式会社　東京都千代田区内幸町1丁目1番6号

㉔代理人　弁理士　森田　寛

明　細　書

1．発明の名称

類似度補正処理方法

2．特許請求の範囲

特徴データと操作データとからなるN個のデータ対が格納されたデータベースから，検索データの特徴を検索キーに，それと類似な特徴を有するデータ対を検索する際に，

2つの特徴データ間の類似度を生成する類似度生成手段と，

前記類似度生成手段を用いて作成した1個の検索データとN個のデータベース内のデータ対とのN次元類似度ベクトルを記憶する検索用類似度ベクトル記憶手段と，

前記類似度生成手段を用いて作成したデータベース内のN個のデータ対相互間のN×N次元類似度ベクトルを記憶する学習用類似度ベクトル記憶手段と，

データベース内のデータ対相互間のN×N次元序列ベクトルを生成する序列生成手段と，

前記N×N次元序列ベクトルを記憶する学習用序列ベクトル記憶手段と，

1個のN次元検索用類似度ベクトル，あるいは1個のN次元学習用類似度ベクトルを入力とし，1個のN次元序列ベクトルを出力とする神経回路網とを設け，

検索に先だって，予め前記学習用類似度ベクトルのうち1個を神経回路網に入力し，教師信号を上記学習用類似度ベクトルに対応した1個の学習用序列ベクトルとし，それぞれ学習用に蓄積されたN個のベクトルについて学習を繰り返すことにより，神経回路網の最適化を行い，

検索データが与えられた時は，検索用類似度ベクトルを前記学習済みの神経回路網へ入力することにより，実際に有益なデータの序列を示す序列ベクトルを得るようにした

ことを特徴とする類似度補正処理方法．

3．発明の詳細な説明

〔産業上の利用分野〕

本発明は，ある構造を持った特徴を検索キーにして，それに類似なデータをデータベースから検索する（以下類似検索）というシステムにおいて，特徴間の構造的類似関係によってのみ求められる類似度に基づいて類似検索するのではなく，経験的な検索傾向による類似度の補正を行い，より実用的な類似なデータ検索を可能にするようにした類似度補正処理方法に関するものである。

例えば類推を用いた化学物質設計支援システムにおいて類似検索を応用した場合，化学物質の反応事例を反応後の化学構造を特徴データとし，反応前の化学構造及び反応条件を操作データとして，多数格納された化学物質反応データベースから，設計対象とする化学物質の化学構造を特徴とする検索キーにして，それと類似した化学物質の反応データをデータベースから検索する必要がある。その場合の設計の良否は，より実用的なデータを検索することである。しかし構造特徴のような一面的な類似性だけにより，類似なデータを検索することは必ずしも，設計に携わる利用者にとって実用的なデータとは限らない。本発明はこのような類推を用いた設計支援システム等に必要な類似検索に関して，検索側のデータ特徴とデータベース内の特徴データとの構造的な最大類似性のみを考慮して，類似データ対を検索するのでは無く，検索されようとしているデータ対の利用頻度や重要性等の経験的な検索傾向をも考慮した類似検索を可能にし，より実用性の高い設計支援システム等を実現可能とすることを狙いとしている。

〔従来の技術〕

初期の類似検索では，特徴の構造的類似性を部分極大マッチング等により，定量的な類似度で表現し，この類似度の序列でデータ検索を行う方法が採用されてきた。さらに，改良方法として，構造的な類似度の序列が実際に有用なデータの順序を保証しているわけではないことから，個々のデータに重み係数を付与し，これにより類似度を補正し，この補正された類似度の序列から，より実用性の高いデータを検索することが考えられてきた。さらに本発明との係わりでは，この重み係数を，データベースから帰納的に決定する方法が提案されてきた（情報処理学会第40回（平成2年度前期）全国大会講演論文集(1) 4D－6「重みを利用した類似化学反応の検索」による）。この従来技術における重みの帰納法的決定方法とは，データベースから単一のデータを引き抜き，そのデータとその他のデータとの間の類似度値の序列を求め，さらにそのデータとその他のデータとの間の実際の有用な序列を類推シミュレーションによって求め，両者の序列間の差分を検出し，最初に求められた類似度序列のうち過小に見積られた類似度を大きくするように，対応するデータに付けられた重みを単位量だけ増加することを，全てのデータについて繰り返し行う方法と考えられる。

〔発明が解決しようとする課題〕

上記の従来技術方式における重み決定方法は，事例に基づいた一種の強化学習を行うのであるが，問題点としては

① 修正が関係するデータに付与された重みに，単位量でしか行われないこと，

② 重み修正が増加方向にしかなされないことのために，必ずしも重み決定の収束保証がないこと

があげられ，これが類似検索の精度を低下させていると考えられる。

本発明は上記従来技術の問題点に関し，①については，関係するデータの重みだけでなく全データの重みに対して，可変量で修正するようにしたこと，また②については，重みの修正を増加方向だけでなく，減少方向にも行うようにしたことであり，これにより重み決定の収束性を向上するとともに，類似度補正の精度を向上することを目的としている。

〔課題を解決するための手段〕

本発明は類似検索における構造的類似度の序列

を実際に有用な類似度の序列にマッピングするための類似度補正に関するものである。これは，データベース中の全データ数（N個）を次元数とするN次元の類似度ベクトルを同じN次元の序列ベクトルに変換することを意味する。この変換機能は，変換事例が類似度ベクトルと序列ベクトルとして獲得できる場合には，事例学習によって構築することができる。本発明では，事例学習の手段として，N×Nの神経回路網を利用することとし，学習に必要な事例については，データベース中のデータから獲得することとする。つまり，データベースからデータを1つ抜き出し，そのデータとデータベース中のN個のデータとのそれぞれの類似度を計算し，それをN次元ベクトル化したものを入力信号，またそのデータとN個のデータとのそれぞれの序列を類推によって推定し，それをN次元ベクトル化したものを教師信号とし，これらを組にして1つの事例データとする。従って，N個のデータについてN組の事例データが得られ，学習段階ではこれらを繰り返し用いることとする。

〔作　用〕

データベースのデータを用いて学習するために同じデータを使った神経回路網による想起が可能であるばかりでなく，データベースに無いデータについても近似的な連想機能が利用できる。つまり，新規のデータに対して，全データとの計算されたN次元類似度ベクトルを入力とした時に，出力にはN次元の実際に有効な序列ベクトルが得られる。

前述のことは神経回路網を使用することにより，従来技術において個々のデータに付与していた重みは神経回路網内部のユニット間の重みに置き換えられ，重みによる類似度補正を神経回路網内のユニット間の重み計算とユニット内の伝達関数の働きによって達成したことであり，さらに収束性の保証されたバックプロパゲーション等の神経回路網を利用した事例学習により，より正確な類似度補正を可能にしたことである。

〔実施例〕

本発明に係わる機能構成について第1図に実施例を示す。

第1図においては，検索用特徴データ1，データベース2，序列変換部3，及び序列データ4から構成される。さらに，第2図は序列変換部の内部構成を示しており，1，2，3，4は第1図に対応し，類似度生成部3－1，序列生成部3－2，検索用類似度ベクトル記憶部3－3，学習用類似度ベクトル記憶部3－4，学習用序列ベクトル記憶部3－5，及び神経回路網3－6，学習機構3－7，及び検索用序列ベクトル記憶部3－8から構成される。

なお，検索用特徴データはデータベース内データ対の特徴データと同一形式であり，検索用類似度ベクトルは学習用類似度ベクトルの1エントリと同一形式であり，検索用序列ベクトルは学習用序列ベクトルの1エントリと同一形式である。

序列データ4は検索用序列ベクトルのサブセットであり，実際に必要なデータ対数だけを提示するために設けられる。但し，この機構については本発明の請求範囲外のことであり詳細は省く。

本発明の動作について説明する。動作は大きく学習フェーズと検索フェーズとから構成され，学習フェーズが検索フェーズに先行する。即ち初期の神経回路網3－6はランダムに初期化されており，実際の検索に先だっては適切な動作が可能なように事例からの神経回路網3－6の重み設定が必要である。これを学習といい，検索段階に先だって行われる必要がある。

以降，学習段階，検索段階の順に実施例を説明する。第8図は本発明の実施例概略機能フローチャートを表す。

（学習段階）

データベース2は第3図図示のように特徴データCiと操作データOiとの対からなるN個のデータ対Di（i＝1～N）が格納されているものとする。学習段階では，先ず学習用類似度ベクトル記憶部3－4と学習用序列ベクトル記憶部3－5とを準備する。学習用類似度ベクトル記憶部3

－4はデータベース2内のN個のデータ対Di（i＝1～N）相互間の類似度sij（i＝1～N，j＝1～N）を類似度生成部3－1により計算し，第4図に示すようにN×Nベクトルとして記憶したものである。また学習用序列ベクトル記憶部3－5は学習用類似度ベクトル記憶部3－4に対応して，データベース内のN個のデータ対Di（i＝1～N）相互間の序列eij（i＝1～N，j＝1～N）を序列生成部3－2により計算し，第5図に示すようにN×Nベクトルとして記憶したものである。類似度生成部3－1は2つの特徴データCp，Cq間の類似度spqを生成するものであり，化学構造図のように特徴データがグラフ構造式で記述されておれば，グラフ間の最大一致度を類似度とすることができる。さもなければ2つの特徴データからそれぞれ同一のカテゴリで特徴表現しベクトル化した2つの特徴ベクトル間での一致数を類似度とすることもできる。いずれにしても類似度は対象とする問題領域毎に適切に作成されたものを利用するものとする。

序列生成部3－2は2つのデータ対Dp，Dq間の序列度を生成するものであり，化学反応のような構造変化に対してはデータベース自身から類推シュミレーションを利用して生成することができる。即ちデータ対Dpについて特徴データCp，特徴データCq，及び操作データOqを利用して操作データOpを類推する類推手段を有することが前提である。1つのデータ対Di内の特徴データCiについてデータベース内の全てのデータ対Djとの間で操作データを類推すると，N個の操作データOij（j＝1～N）が得られる。このN個の操作データOij（j＝1～N）をデータ対Di内の既知である操作データOiと比較し，類似度の大きいものから順番に1から最大Nまでの序列を付ける。このデータ対Diのデータ対Djに対する序列をeij（j＝1～N）とし，全てのi（i＝1～N）について序列eij（i＝1～N，j＝1～N）を求めればN×Nのベクトルが生成される。但し，類推が不能な場合は序列の値は'0'とする。なお，この序列化に用いる類推については対象とする問題に依存するものである。以下化学反応を例とした場合の類似度生成，及び序列生成について説明する。

第6図は類似度生成部3－1の働きを示したものである。2つの特徴データCp，Cqが示されている。特徴データCp，Cqは化学構造を特徴インデックスとして捉えた表現が用いられており，類似度spqは特徴データ内のインデックス間での一致度の総得点としている。あるいはここで各インデックス間の一致度に適当に重みを付加した加重加算による総得点とすることもできる。即ち，特徴をCp＝（cp1，cp2，…cpk）とCq＝（cq1，cq2，…cqk）とした場合，類似度spqは以下のように計算される。

$$spq=\sum_{k} a_k \{cpk, cqk\}$$

但し，{cpk，cqk}；特徴インデックスk間の一致度

a_k；特徴インデックスkの重み係数

例えば，第6図では2種の芳香族単環化学物質間の類似度を化学構造特徴インデックスを適当に設定した場合に，類似度が22と計算されることを示している。このような計算手法を用いて，あるデータ対Di内の特徴データCiについて他のデータ対Dj内の特徴データCjとの類似度sijを求め，N次元の類似度ベクトルSi＝（si1，si2，…siN）を作成する。さらに全ての特徴データCiについて同様にN個のN次元類似度ベクトルSi（i＝1～N）が生成でき，これを学習のために使用する学習用類似度ベクトル記憶部3－4に格納する。

次に序列生成部3－2の働きについて説明する。

化学反応は化学構造の変化であり，その変化の前後がそれぞれデータ対内の特徴データC，及び操作データOに格納されているものとすると，同種の化学変化を持つ複数のデータ対相互では化学構造変化が類似しており，一方の変化前あるいは変化後の化学構造が無くても類推により生成可能である場合が多い。この類推の手法があるものとすると，2つのデータ対Dp，Dqについて一方

の特徴データＣｐのみと他方のデータ対（Ｃｑ，Ｏｑ）とから操作データＯｐｑを類推し，これが既知の操作データＯｐと類似している場合，この２つのデータ対ＤｐとＤｑは類似していると考えられる。この類似性は構造的な類似度ａｐｑとして容易に生成可能であり，先の類似度生成機構を使用しても良い。つまり，ａｐｑは以下のように定義できる。

ａｐｑ＝〔Ｏｐ，Ｏｐｑ〕＝〔Ｏｐ，〔Ｃｐ，Ｃｑ，Ｏｑ〕〕

但し，〔Ｏｐ，Ｏｐｑ〕　；操作データ間の類似度

Ｏｐｑ　＝〔Ｃｐ，Ｃｑ，Ｏｑ〕

〔Ｃｐ，Ｃｑ，Ｏｑ〕；Ｏｐｑ の類推

１つのデータ対Ｄｉについて他のデータ対Ｄｊ（ｊ＝１～Ｎ）との類似度ａｉｊを求め，１つの類似度ベクトルＡｉ＝（ａｐ１，ａｐ２，…ａｐＮ）が求められるが，本発明ではこれを１からＮの序列値で序列化したものを序列ベクトルＥｉ＝（ｅｉ１，ｅｉ２，…ｅｉＮ）とする。つまり各データ対Ｄｉについて，１個のＮ次元序列ベクトルＥｉが生成でき，全データ対ＤｉについてはＮ×Ｎ次元の序列ベクトルＥｉ（ｉ＝１～Ｎ）が生成でき，これを学習のために使用する学習用序列ベクトル記憶部３－５に格納する。なお，類推時に類推が不能な場合がある。例えば，２つのデータ対Ｄｉ，Ｄｊ間に同一化学反応を持つことができない場合，即ち反応中心を共有できない場合は類推するまでもなく，類似度ａｉｊは‘０’であり，序列値ｅｉｊも‘０’とみなされる。

第７図は４つのデータ対Ｄｐ，Ｄｑ，Ｄｒ，及びＤｖについてＤｐとＤｑ，Ｄｒ，及びＤｖとの序列生成例を示したものである。なお，ここではＤｖのように反応中心を共有できない場合や，Ｄｒのように類推によって生成したＯｐｒとＯｐが一致しない場合には類似度ａｐｒ，ａｐｖは‘０’とし，Ｄｑの場合のように類推によって生成したＯｐｑとＯｐが一致する場合にはａｐｑは‘１’とした。この場合には類似度ａｐｊ（ｊ＝ｑ，ｒ，ｖ）はそのまま序列ｅｐｊ（ｊ＝ｑ，ｒ，ｖ）として使用される。

次に，神経回路網の学習について説明する。神経回路網３－６を初期化後，学習用類似度ベクトル記憶部３－４からｉ＝１の前期類似度ベクトルＡｉが神経回路網３－６の入力に加えられ，学習機構３－７へは教師信号として学習用序列ベクトル記憶部３－５からｉ＝１の序列ベクトルＥｉが与えられる。学習機構では神経回路網の出力である序列ベクトルＥｉと学習機構に与えられた予め計算済みの序列ベクトルＥｉとから誤差量を計算する。誤差量はバックプロパゲーション学習アルゴリズムで最も一般的な二乗誤差等が用いられる。学習機構３－７ではこの誤差の変化を基にバックプロパゲーション学習を行い，神経回路網の重み等を修正する。引き続き，学習サイクルをｉ＝２からｉ＝Ｎまで行い，さらにこれを誤差が一定の値まで降下するまで同一学習用類似度ベクトル記憶部３－４，及び学習用序列ベクトル記憶部３－５内のそれぞれ類似度ベクトル，及び序列ベクトルを複数回用いて学習サイクルを繰り返す。以上により，出力の序列ベクトルと期待される序列ベクトルとの誤差がある一定量以下になった場合に神経回路網３－６の学習は終了したものとみなし，学習フェーズを終了する。

（検索段階）

検索フェーズではデータベース内の特徴データと同じ形式の検索データＣｔが与えられたとした場合には，類似度生成部３－１の働きにより，検索データＣｔとデータベース内の全特徴データＣＩ（ｉ－１～Ｎ）間で類似度Ｓｔｉが計算されｌつのＮ次元類似度ベクトルＳｔ＝（ｓｔ１，ｓｔ２，…ｓｔＮ）が作成され，検索用類似度ベクトル記憶部３－３に記憶される。この類似度ベクトルを学習済みの神経回路網３－６に入力すると，対応する序列ベクトルＥｔ＝（ｅｔ１，ｅｔ２，…ｅｔＮ）を検索用序列ベクトル記憶部３－８に出力する。この序列ベクトルにおいて，表示される序列が実際に検索されるべき類似なデータ対の序列を示しており，前段の学習が充分精練されているならば，この序列は経験的な検索傾向を反映するはずであり，実際には，そのうちの上位いくつかを序列データ４として出力する。この序列データが示すデータベース内のデータ対を利用することによって，

本発明の目的とするより有効な設計が可能となる。

〔発明の効果〕

以上の説明から明らかになるように，本発明によれば，ある構造を持った特徴を検索キーにして，それに類似なデータをデータベースから類似検索する場合，特徴間の構造的類似関係によってのみ類似なデータを検索するのではなく，経験的な検索傾向をも反映した類似なデータ検索を行うことが可能になる。上記実施例では化学物質設計支援の場合について示したが，他に例えば，ソフトウェア設計支援システム等において類推手法によるプログラムモジュールの再利用を行う場合，単にプログラムの特徴が似ているということだけで類似なプログラムを検索するのではなく，プログラムの利用度，重要性，経験的な有効性等を反映した類似度に補正でき，その結果，より利用価値の高いプログラムを見つけ出すことができる等，このような類似検索を必要とするシステムは多く考えられ，知識ベースから類似な知識を検索して類推を行う他の一般的な知識ベースシステムについても同様な効果が期待できるという著しい工業的効果がある。

4．図面の簡単な説明

第1図は類似度補正システムの機能構成，

第2図は序列変換部の機能構成，

第3図はデータベース内のデータ対構成，

第4図は学習用類似度ベクトル記憶部構成，

第5図は学習用序列ベクトル記憶部構成，

第6図は類似度生成部の働き，

第7図は序列生成部の働き，

第8図は類似度補正の概略機能フローである。

図中において，

1…検索用特徴データ

2…データベース

3…序列変換部

3－1…類似度生成部

3－2…序列生成部

3－3…検索用類似度ベクトル記憶部

3－4…学習用類似度ベクトル記憶部

3－5…学習用序列ベクトル記憶部

3－6…神経回路網

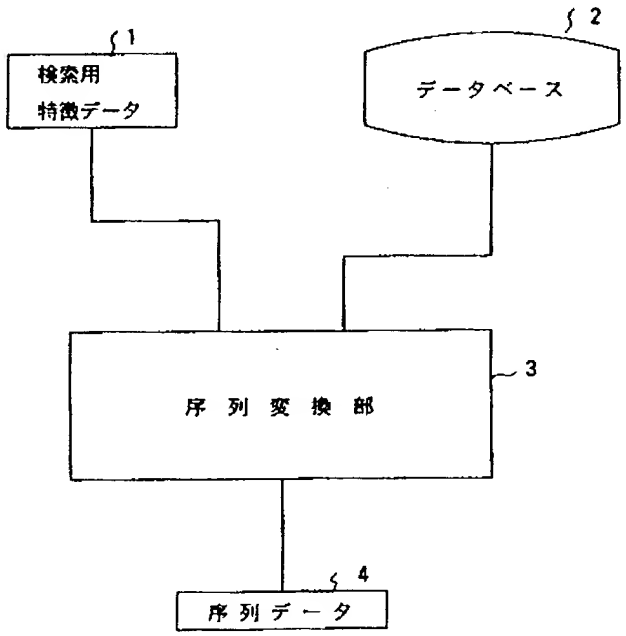
3－7…学習機構，

3－8…検索用序列ベクトル記憶部

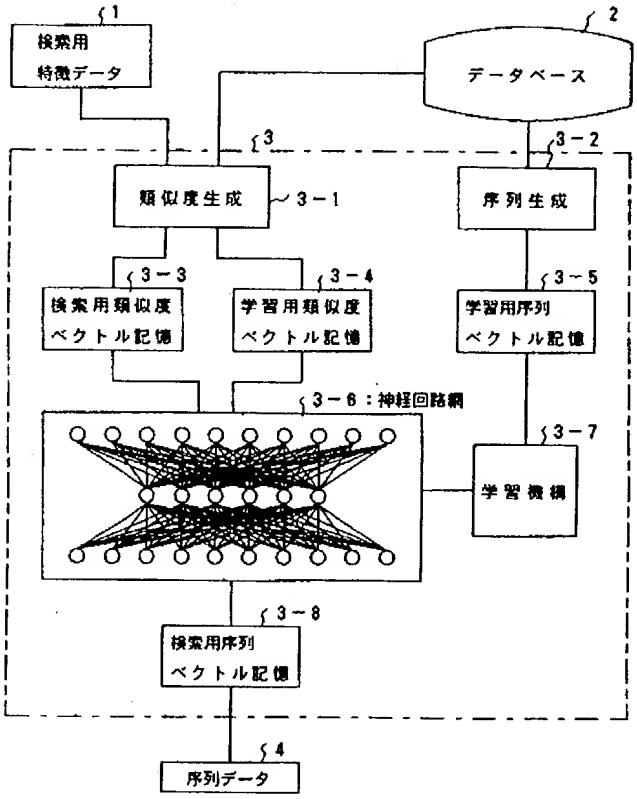
4…序列データ

特許出願人　日本電信電話株式会社

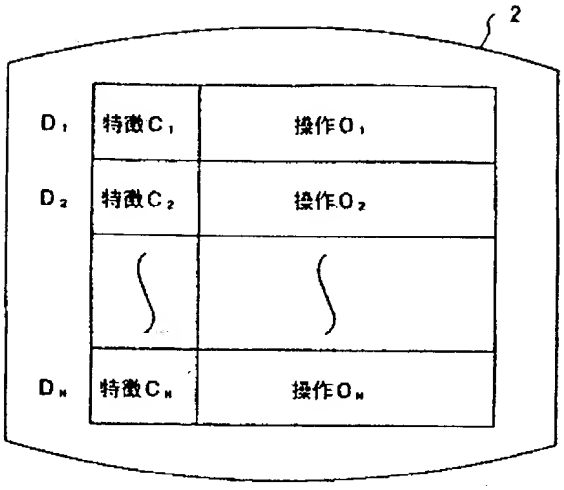
代　理　人　弁理士　森　田　　寛



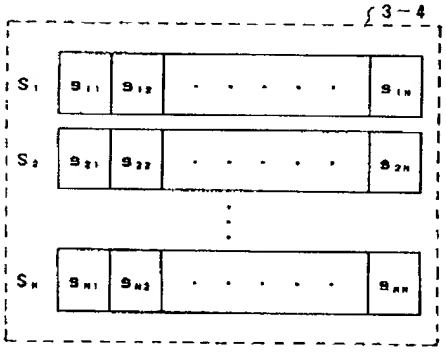
第1図



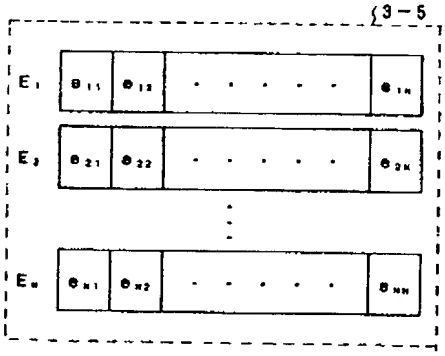
第 2 図



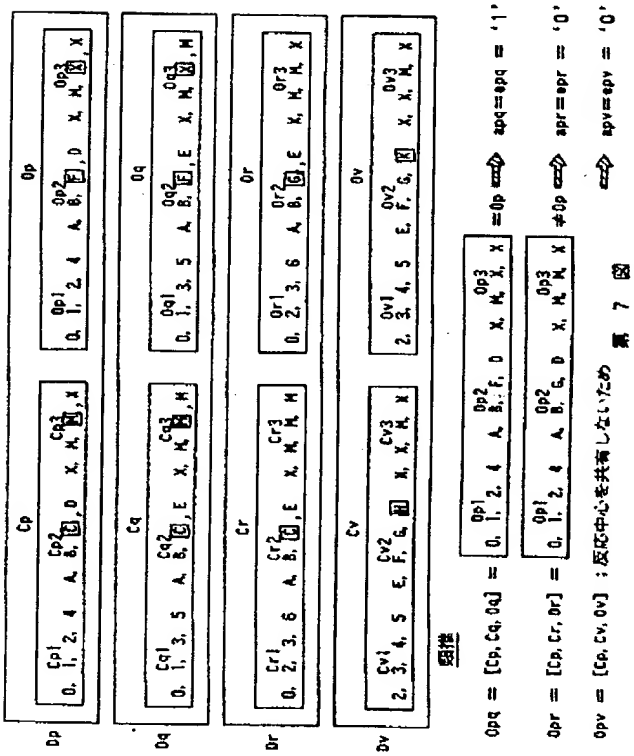
第 3 図



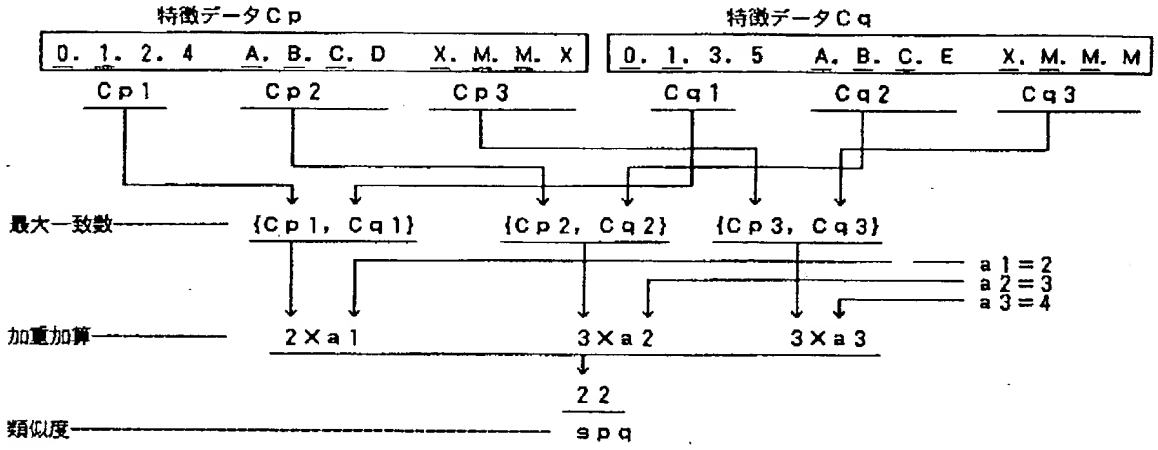
第 4 図



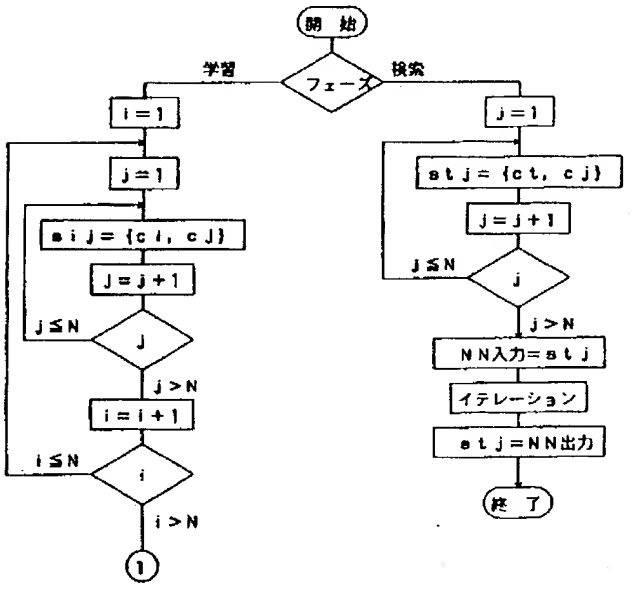
第 5 図



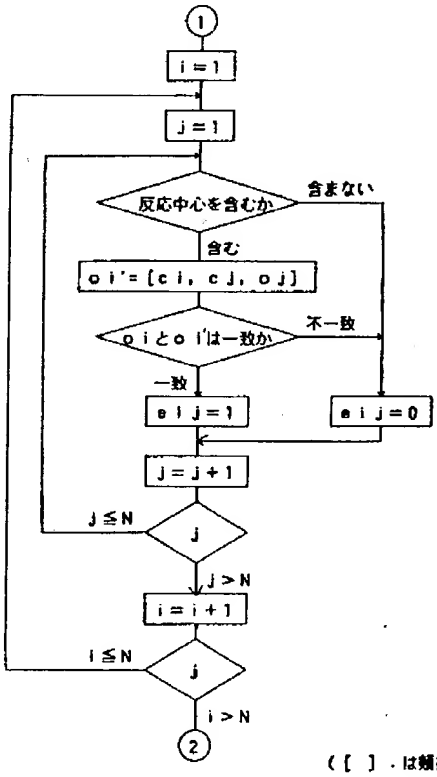
第 7 図



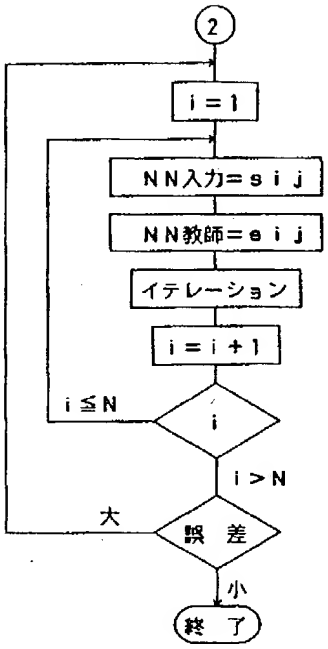
第 6 図



第 8 図(I)



第 8 図(II)



第 8 図(III)